Multi-Feature Extraction Method of Power Customer's Portrait based on Knowledge Map and Label Extraction

Wentao Liu*, Liang Ji
Big Data Center of State Grid Corporation Limited
Beijing 100053, China

***Abstract*—In order to realize the visualization of power customer characteristics and better provide power services for power customers, a multi-feature extraction method of power customer's portrait based on knowledge map and label extraction is studied. The power customer's portrait construction model is designed, which uses the knowledge map construction link to collect the power customer related data from the power system official website and database, and clean and convert the data; In the multi-feature analysis section, natural language processing technology is used to analyze the characteristics of power customers through Chinese word segmentation, vocabulary weight determination and emotion calculation; Based on the feature analysis results, the portrait label is extracted to generate the power customer's portrait. The power customer's portrait is used to realize the application of power customer's feature visualization, power customer recommendation, power customer evaluation and so on. The experimental results show that this method can effectively construct the knowledge map of power customers, accurately extract the characteristics of power customers, generate labels, and realize the visualization of power customer's portraits.**

***Keywords*—Knowledge map; label extraction; power customer's portrait; multi-feature extraction; natural language processing; feature visualization**

I. Introduction

At present, China's power grid has retained nearly 500 million customer data, which come from data collected by power grid energy meters, customer telephone service system data, power grid management data, etc. [1]. Based on these data, using big data analysis technology to analyze power customers, we can build a multi-level and multi-dimensional power customer's portrait [2]. The customer portrait can help relevant personnel quickly and accurately identify and recognize customers, and formulate targeted, refined and personalized service plans by quantifying sensitivity, so as to improve service quality and service efficiency [3].

The current research on portraits can be roughly divided into two categories according to the different objects of the portraits: one is the user portraits of scientific and technological experts, library users, website users and other characters, which are mainly used to recommend content to the research objects and improve the experience of the research objects; The other is the portrait of non-human things such as cities and books [4], which is mainly used to provide information services for groups other than the research object based on the portrait of things. Among them, the research on user portrait is relatively more extensive, and has been extended to various fields for different application scenarios. A lot of research has been done on the generation methods of user portraits at home and abroad. Wang and Zhu et al. used user portraits sharing representative opinions to predict emotion, and considered the impact of different characteristics of user groups on emotion analysis from three aspects: attribute characteristics, interest characteristics and emotional expertise [5]. Guo and Wei et al. used user portraits to automatically obtain user behavior data in web server logs in a big data environment by using a collection system; used PrefixSpan algorithm to build user portrait model, establish user behavior feature labels through frequent sequence mining, and realize abnormal feature extraction [6]. Chicaiza et al. used knowledge maps to build user portraits, and designed a recommendation system based on user portraits to achieve a good recommendation function [7].When the above three methods process massive power customer data, it is difficult to achieve rapid data extraction, and to provide better power service, and the application value is not high. Based on the previous research results, the multi-feature extraction method of power customer's portrait based on knowledge map and label extraction is studied. Knowledge atlas aims to describe the entities in the objective world we know and the direct relationships between entities in a structured form. It provides a better way to summarize and manage massive data in the Internet. The application of knowledge map is a good thing for the semantic search of the Internet. It plays an absolute role in promoting the development of semantic search. At the same time, it also shows its special and amazing ability in intelligent question answering. In the era of big data, knowledge map will have its place. Big data is the foundation and knowledge map is the tool. Combined with deep learning, it will develop better and better. Based on the analysis of user characteristics, tag extraction abstracts user portraits into user-friendly or computer-readable labels. The label has generality and condenses the key information in user characteristics. The content of user portrait labels is diverse, which can be words, phrases or concepts. The visualization methods of labels can be vectors, description charts and label clouds. Through the full analysis of user features and the accurate extraction of labels, we can achieve a deeper characterization of user portraits. By accurately constructing the power customer's

*Corresponding Author.

portrait, we can better provide power services for power customers.

II. Materials and Methods

A. Construction of Power Customer's Portrait

In the context of big data, single data cannot fully reveal the characteristics of things, and the decision support services are insufficient and incomplete. Multi-source big data can describe things from different perspectives. Data complement and cross verify each other, breaking the "data island" [8]. Based on the understanding and elaboration of domestic and foreign scholars on the concept of power customer's portrait, it is considered that power customer's portrait refers to a model that can describe and characterize the features of power customers from different dimensions after abstracting and summarizing the basic statistical characteristics of power customers, power category characteristics, contact preference characteristics, power grid sensitive type characteristics and other characteristics from multi-source big data. On this basis, a model based on power customer's portrait is proposed, as shown in Fig. 1.

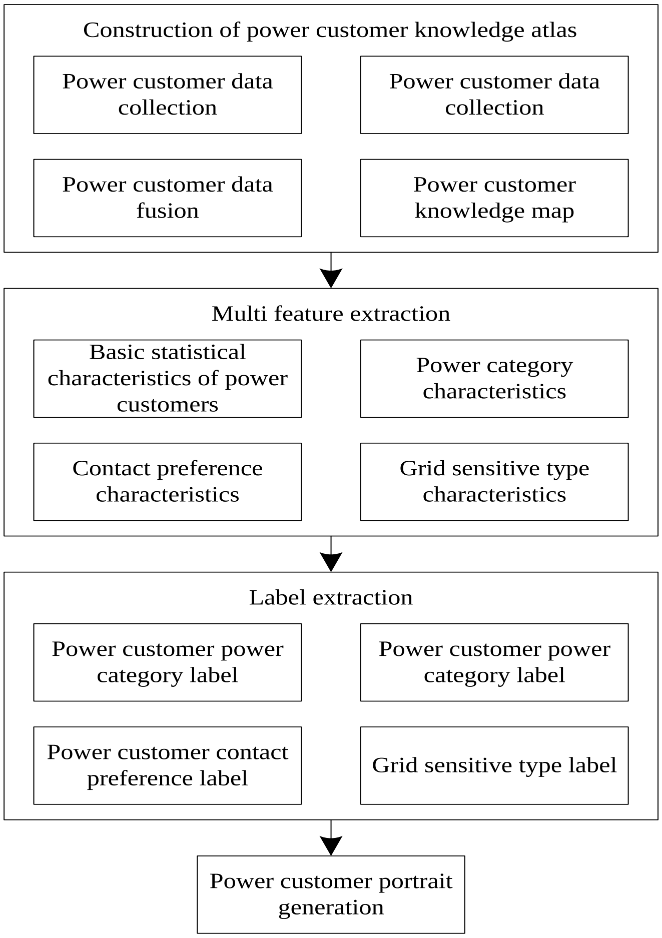
Construction of power customer knowledge atlas
Power customer data collection
Power customer data collection
Power customer data fusion
Power customer knowledge map
Multi feature extraction
Basic statistical characteristics of power customers
Power category characteristics
Contact preference characteristics
Grid sensitive type characteristics
Label extraction
Power customer power category label
Power customer power category label
Power customer contact preference label
Grid sensitive type label
Power customer portrait generation

Fig. 1. Power Customer's Portrait Construction Model.

The power customer's portrait construction model can be roughly divided into three links: knowledge map construction, multi-feature analysis and portrait label extraction.

The knowledge map construction link collects relevant data of power customers from the power system official website, encyclopedia website and literature database; Data cleaning, data specification, data integration, data conversion and other data processing work are carried out for the relevant data of power customers. In the step of data fusion, the relevant data of power customers obtained after processing are stored according to the basic statistical characteristics of power customers, the characteristics of power categories, the characteristics of good contacts, the characteristics of power grid sensitive types and other categories to realize data processing; Multi-feature analysis links analyze the characteristics of power customers; Based on the feature analysis results, the portrait label is extracted to generate the power customer's portrait. The power customer's portrait is used to realize the application of power customer's feature visualization, power customer's recommendation, power customer's evaluation and so on.

B. Construction of Power Customer's Knowledge Map

The data extraction and processing of power customers in the process of building the knowledge map of power customers is to collect the relevant data of power customers from the power system website, encyclopedia website and relevant structured knowledge base, and build the original tourism route database.

The perfection of the construction of power customer's knowledge map directly affects the quality of the generation results of power customer's portraits [9]. Therefore, it is necessary to collect relevant data in the official website of the power system, encyclopedia websites and literature databases, and analyze the collected data through big data technology to determine that it meets the construction standard of knowledge map. On this basis, the collected power customer's related data is analyzed with the current existing data [10], and the data that meets the threshold standard is defined as valid data, which is stored in the original database.

Fig. 2 shows the construction process of power customer's knowledge map. The construction process of power customer's knowledge map can be roughly divided into four links:

1) *Power customer data collection:* The main function of this link is to collect relevant data of power customer's entities in various websites, and realize noise elimination through the preprocessing process in the data crawling process. On this basis, based on the differences of data types (structured, semi-structured and unstructured), the collected data are stored and processed, and the power customer entity database is constructed.

2) *Power customer data extraction:* The main function of this link is to use the data extraction model to extract different data needed to build the power customer's knowledge map from all kinds of data collected.

3) *Power customer data fusion:* The main function of this link is to use thesaurus to fuse the extracted data, and in this process, to complete the elimination of entity / attribute ambiguity and synonymous relationship merging [11].

4) Construction of power customer knowledge map: The main function of this link is to use the fused data to generate entity / attribute / relationship triples, so as to finally complete the construction of power customer knowledge map.

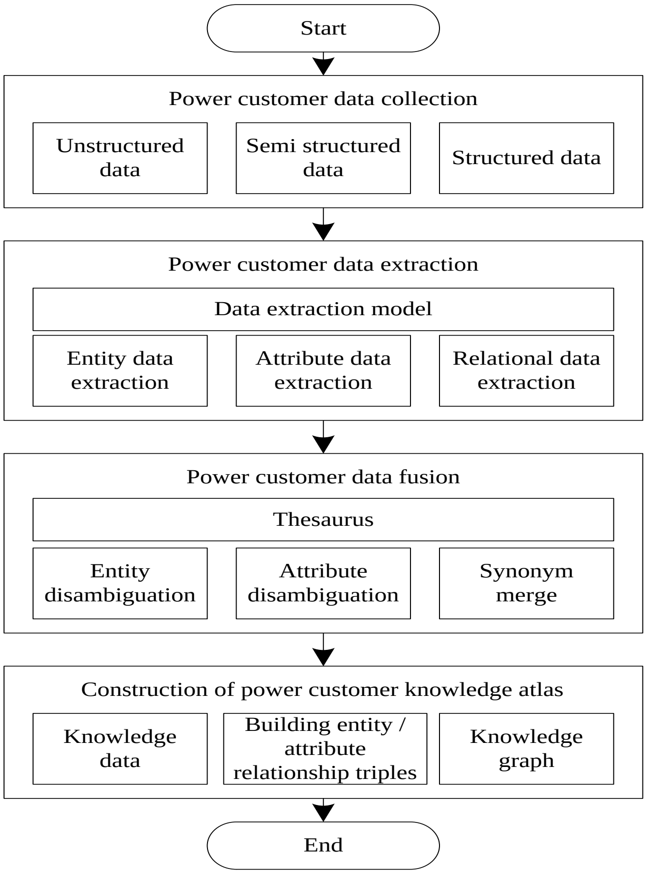
Start
Power customer data collection
Unstructured data
Semi structured data
Structured data
Power customer data extraction
Data extraction model
Entity data extraction
Attribute data extraction
Relational data extraction
Power customer data fusion
Thesaurus
Entity disambiguation
Attribute disambiguation
Synonym merge
Construction of power customer knowledge atlas
Knowledge data
Building entity / attribute relationship triples
Knowledge graph
End

Fig. 2. Construction Process of Power Customer Knowledge Map.

In the data extraction process, a timestamp incremental data extraction model based on variable window is used, which depends on the given time window for incremental data extraction. It is normal to assume that exceptions occur in the process of data extraction, and the occurrence of exceptions often leads to the inconsistency between the target table data and the source system data. To solve this problem, before incremental data extraction, it firstly gives two time windows (source time window and target time window, which are respectively used to extract data from various websites and power customer entity databases), then uses the given time window to extract corresponding data from various websites and power customer entity databases, and finally de-duplicates these data to produce the final result set. Considering the impact of abnormal occurrence on the consistency of power customer data in the process of power customer data extraction [12], it is necessary to use the data maintenance lead time to make minor adjustments to the given time window before power customer data extraction. Here, the formal definition of timestamp incremental data extraction model with variable window is given:

$$M=\left(S_R,d_s,gmt,\Delta t,T_W,O_P,clean,d_u\right) \tag{1}$$

In equation (1), S_R is an n-tuple, $S=\left(s_1,s_2,\cdots,s_n\right)$ represents n data sources; d_s refers to the information database table of power customer, which requires an attribute column that marks the storage time of records; gmt represents the maximum value of the warehousing time of the data obtained from table d_s; Δt refers to the lead time of power user's data maintenance, which is used to adjust the size of the time window. T_W represents the time window information required for data extraction, T_W can be represented by a binary $T_W=\left(T_{W,S},T_{W,T}\right)$, where $T_{W,S}$ represents the source time window and $T_{W,T}$ represents the target time window. In model M, the determination of $T_{W,S}$ and $T_{W,T}$ in T_W is very important to ensure the data consistency between the source system and the power customer's information database, especially when the data extraction process is started again after an exception occurs in the data extraction process, the determination of $T_{W,S}$, in T_W is directly related to whether the data that cannot be extracted due to the exception can be extracted. O_P refers to the operation defined on the data source, O_P is also a multivariate group, and the dimension of tuples is related to the operation defined on the data source, such as $O_P=\left(INNERJOIN_{s_1-s_2},\cdots\right)$, which means that the power user data extracted from data source s_1 and data source s_2 will be internally connected. $clean$ represents a simple cleaning operation for the power user data extracted from S_R, that is, to remove the redundant time field value (the non-minimum data warehousing time of each record from S_R); d_u means to de duplicate the power user data extracted from S_R and d_s.

The working process of incremental data extraction model based on variable time window is described as follows:

1) Perform gmt operation on the target database table d_s to obtain the maximum value $\max Time$ of the storage time of power user data in d_s;

2) Set the value of the start time node of data extraction in $T_{W,S}$ and $T_{W,T}$ time windows to $\max Time$ respectively, and then adjust the time windows $T_{W,S}$ and $T_{W,T}$ according to the given data maintenance lead time to obtain $T_{W,S}{}'$ and $T_{W,T}{}'$;

3) Take $T_{W,S}{}'$ as the new time window, extract the data from the power user data source $S_{R,i}$, and operate the extracted power user data O_P, and then make $clean$ operation to remove the redundant time field values in the data to obtain the result set $tempRS\_S_R$;

4) Then take $T_{W,T}{}'$ as the target time window, extract the power user data from d_s, and get the result set $tempRS\_d_s$;

5) Perform d_u operation on $tempRS\_S_R$ and $tempRS\_d_s$ to obtain the final result set;

6) Load the final result set into d_s.

C. Multi-feature Extraction based on Natural Language Processing Technology

Multi-feature extraction of power grid sensitive customer's portrait under natural language technology, that is, using the methods of word segmentation, word vector conversion and word weight calculation in natural language processing technology to extract the features in the power grid sensitive customer's portrait label.

1) Chinese word segmentation: Letting the computer understand the label text information is the basis of multi-feature extraction of sensitive customer's portrait in power grid. There are a large number of Chinese words in the label text translated from Chinese. When using computer tools to comprehensively analyze the text information of portrait labels, the more common statistical word segmentation method is usually used [13] to preprocess the text information of portrait labels of power grid sensitive customers.

The N-ary grammar model in the statistical analysis method is selected. The conditions for the establishment of this model are: the production of the N-th word is affected by the first N-1-th word; There is no correlation with other words except this word and the N-1-th word; The probability of a whole sentence in the label text is the product of the production rates of different words in the sentence. Using the N-ary grammar model to judge the scientificity of the word segmentation plan is based on the probability value of the sequence of N words [14]. For example, suppose a sentence Y in the label text of power grid sensitive customer's portrait has two word segmentation plans, which are defined as Y_1 and Y_2 respectively, and the detailed process of word segmentation processing using the N-ary grammar model is as follows:

a) If the word order in Y is $C^{(1)}, C^{(2)}, C^{(3)}, \cdots, C^{(n)}$, it is expressed as $Y_1 = \left\{ C_1^{(1)}, C_1^{(2)}, C_1^{(3)}, \cdots, C_1^{(n)} \right\}$ and $Y_2 = \left\{ C_2^{(1)}, C_2^{(2)}, C_2^{(3)}, \cdots, C_2^{(n)} \right\}$ respectively.

b) The joint probability distribution expression of Y_1 and Y_2 is:

$$P(Y_m) = P\left(R_m^{(1)}\right) \times P\left(C_m^{(2)} \middle| C_m^{(1)}\right) \times P\left(C_m^{(3)} \middle| C_m^{(1)} C_m^{(2)}\right) \times \cdots \times P\left(C_m^{(n)} \middle| C_m^{(1)} C_m^{(2)} \cdots C_m^{(n-1)}\right) \tag{2}$$

In equation (2), $P\left(C_m^{(n)} \middle| C_m^{(1)} C_m^{(2)} \cdots C_m^{(n-1)}\right)$, $P\left(C_m^{(n)}\right)$ and $P\left(C_m^{(n)} \middle| C_m^{(n-1)}\right)$ $C_m^{(n)}$ respectively represent the joint probability distribution of sentence Y_m, the probability of word and the basis of word $C_m^{(n-1)}$. Through the public corpus, we can determine the probability of corresponding words, and then we can determine $P\left(C_m^{(n)}\right)$ and $P\left(C_m^{(n)} \middle| C_m^{(n-1)}\right)$.

c) The $P(Y_1)$ probability value is compared with the $P(Y_2)$ probability value, and the higher probability value is the correct word segmentation plan of the label text sentence T of power customer's portrait. After the word segmentation plan is determined, word vector conversion is performed on the label text.

2) Determination of vocabulary weight: Different words form sentences, and different sentences form the label text of the sensitive customer's portrait in power grid. The importance of each word in different label texts is different, and the importance of the same word in different label texts is also different. Therefore, the determination of vocabulary weight is very important for the extraction of multiple features of the sensitive customer's portrait in power grid. Statistical knowledge is used to determine the vocabulary weights that have been transformed into vectors in the text of sensitive customer's portrait labels in power grid, that is, the vocabulary weights are determined according to the text statistical information such as word frequency [15]. The above statistical process is based on Shannon's informatics theory: assuming that a certain word in all texts has a high word frequency, its information entropy is small; On the contrary, the lower the word frequency of a word in all texts is, the greater the information entropy is, that is, the inverse relationship between the word frequency and its information entropy.

As a typical information retrieval and data mining weighting technology, the term frequency-inverse document frequency (TF-IDF) algorithm can act on the vocabulary weight determination of the label text of the sensitive customer portrait in power grid [16], which is conducive to extracting the characteristic vocabulary in the label text information and realizing the multi-feature extraction of the power grid sensitive customer's portrait. The calculation process of TF-IDF algorithm is as follows, in which equation (3) and equation (4) calculate word frequency and inverse text frequency index respectively.

$$F_{i,j} = \frac{x_{i,j}}{\sum_{k} x_{k,j}} \times \ell \tag{3}$$

$$DF_i = \log \frac{T}{j : r_i \in t_j} \times \ell \tag{4}$$

$$F - DF = F_{i,j} \times DF_i \times \ell \tag{5}$$

In the above equation, $x_{i,j}$ and $\sum_{k} x_{k,j}$ respectively represent the occurrence frequency of word i in the power customer's portrait label text t_j and the total number of words in t_j; $|T|$ and $\left|\left\{ j : r_i \in t_j \right\}\right|$ respectively represent the number of all tagged texts in the corpus and the number of all tagged texts containing the word r_i in the corpus; ℓ represents the correction factor. Equation (3) is multiplied by equation (4), that is, equation (5) can exclude the vocabulary commonly existing in the text of power customer's portrait labels, and retain the vocabulary that can reflect a certain feature.

3) Emotion calculation: Using the feature words retained after the vocabulary weight is determined, the sentence emotion in the image label text of power grid sensitive customers can be determined. When determining sentence emotion based on vocabulary level, in order to ensure the efficiency of emotion determination, feature dimensionality reduction needs to be implemented, which will cause a large loss of feature words and lead to the deviation of emotion determination results. In order to avoid such problems, the Latent Dirichlet Allocation (LDA) model is used to determine the sentence emotion in the label text of the sensitive customer's portrait in power grid [17].

$T = \left\{ t_1, t_2, \cdots, t_T \right\}$ and $t_i = \left\{ y_1^i, y_2^i, \cdots, y_k^i \right\}$ are used to represent the text set of the power customer's portrait label and the sentence set in the text t_i respectively, so $T = \left\{ y_1, y_2, \cdots, y_N \right\}$ can be used to describe T as the set of all sentences in t_i, where N represents the number of all sentences in D.

The model of T is constructed by using LDA model. After parameter estimation, the distribution of sentences on characteristic topics and the distribution of characteristic topics on vocabulary are obtained respectively [18], and expressed by $\rho\left(\phi \middle| \alpha\right)$ and $\rho\left(w_n \middle| \phi, \partial\right)$, where ϕ, α and ∂ represent vectors, k-dimensional Dirichlet parameters and a $k \times v$ matrix respectively. Based on $\rho\left(\phi \middle| \alpha\right)$ and $\rho\left(w_n \middle| \phi, \partial\right)$, the multi-feature extraction of power customer's portrait is carried out. The specific process is as follows.

4) Density features: describe the maximum number of sentences with emotional words in the sentence, expressed by density (S).

5) Range features: describe the number of words in the sentence, expressed by range (S).

6) Quantitative features: describe the number of emotional words, expressed by the number (S).

7) Polarity features: q and $p_o(\text{q})$ are used to represent emotional words and their polarity respectively. After calculation, it can determine the polarity of the sentence as:

$$p_o\_s(S) = \sum_{i=1}^{count(S)} p_o(q_i) \times \beta \tag{6}$$

Where β is any constant.

8) Degree features: assuming the degree of q is $e_x(q)$, the degree of the sentence after calculation is:

$$e_x\_s(q) = \max_{i=1,count(S)} e_x(q_i) \times \beta \tag{7}$$

Through Chinese word segmentation, vocabulary weight determination and emotion calculation, the basic statistical features of power customers, power category features, contact preference features and power grid sensitive type features are extracted. According to the feature extraction results, the basic statistical labels of power customers' portraits, power category labels, contact preference labels and power grid sensitive type labels are extracted.

D. Label Extraction and Portrait Generation of Power Customer's Portrait

1) Basic statistical label extraction of power customer's portrait.

$$D_e = \langle B_a, E_d, W_o, C_o \rangle \tag{8}$$

Where, B_a refers to the basic information of power customers, including name, gender and date of birth; E_d refers to the education experience of power customers, including education background, graduation college and graduation time; W_o refers to the work experience of power customers; C_o refers to the contact information of power customers, including office phone, mobile phone, email and personal web address.

2) Electricity category label extraction of power customer

$$S_p = \langle s_1, s_2, s_3, \ldots, s_n \rangle \tag{9}$$

Where, s refers to the power category feature word extracted from the power consumption data of power customers.

3) Contact preference label extraction of power customer

$$I_n = \{(t_1, tf_1), (t_2, tf_2), (t_3, tf_3), \ldots, (t_n, tf_n)\} \tag{10}$$

Where, t refers to the keyword in the contact preference of power customers, and tf_n refers to the number of times the preference word appears.

4) Extraction of grid sensitive type label

$$l_a = \langle l_{a,T}, R_d, R_{di} \rangle \tag{11}$$

Where, $l_{a,T}$ is the node label in the grid use partnership of power customer, R_d is the contribution of power customer d in the power grid use partnership, and R_{di} is the weight of the correlation between power customer d and power customer i .

Power customer's portrait is a label combination model of multi-dimensional and multi-level power customer [19]. According to the type of relevant data of power customers, it can define a vector space to represent the power customer's portrait [20]:

$$d = D_e, S_p, I_n, l_a \tag{12}$$

III. Experimental Results

A. Experimental Environment

In order to verify the application effect of the multi-feature extraction method of power customer's portrait based on knowledge map and label extraction studied in this paper in the actual application process, the open data set of 2020 "Customer portrait" competition is taken as the experimental object, which includes the power use data of millions of grid customers in a province in 2019. This paper determines three types of information, a total of 12 fields (Table I) for the experiment.

The experimental environment is as follows: the server and client exist in the same computer, and the computer processor, memory and operating system are (Intel)i9-9900KF core eight core CPU, 8 GB memory and Windows 8 operating system respectively.

TABLE I. Description of Experimental Data

Experimental information category	Number of fields	Content description
Power system work order information	6	Name, gender, contact information, contact address, customer nature, etc
Registration form information	3	Business hall, store, communication reasons, etc
Data acquisition information of power meter	7	High energy consumption type, electricity fee guarantee type, total electricity consumption, paid in amount, payment, etc

B. Construction Results of Power Customer's Knowledge Map

The construction of power customer's knowledge map is the basis of power customer's portrait generation. According to the experimental data, the proposed method is used to build the customer knowledge map. Fig. 3 shows the customer knowledge ontology model built by this method.

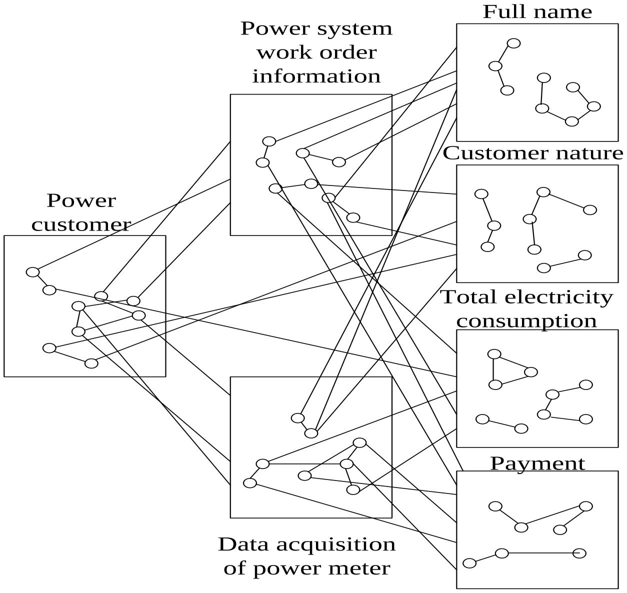

Fig. 3. Experiment Object Knowledge Ontology Model.

The knowledge ontology model of power customer shown in Fig. 3 contains the work order information of the power system of the power customer and the data collection information of the power meter, in which the name, customer nature, total electricity consumption and payment are the low redundancy entity summaries generated by the proposed method. The different kinds of data contained in the body of knowledge come from experimental data.

C. Multi-feature Extraction Test of Power Customer

The multi-feature extraction of power customers is the most critical and time-consuming link in the generation of power customer's portraits. Therefore, the multi-feature extraction test of power customers is mainly analyzed from two aspects: extraction accuracy and extraction efficiency.

In the process of extracting accuracy test, we mainly use accuracy, precision rate, recall rate and F1 value as evaluation indicators to evaluate the effect of experimental feature extraction.

According to the combination of the actual feature extraction results of the experimental object and the extraction results of the proposed method, the real examples, false positive examples, true negative examples and false negative examples are defined. The confusion matrix of the feature extraction results during the experiment is shown in Table II.

TABLE II. CONFUSION MATRIX OF FEATURE EXTRACTION RESULTS

Actual feature extraction results		Positive example	Counterexample
This method extracts the knot	Positive example	True example	False positive cases
	Counterexample	False counterexample	True counterexample

As the most critical and widely used evaluation index in the process of multi-feature extraction of power customers, accuracy can be understood as the proportion of the correct number of samples for feature extraction in the number of text samples of all power customers' portraits, and its expression is:

$$\text{Accuracy}=\frac{\text{ture positive example+false negative example}}{\text{ture positive example+false negative example+ ture negative example+false positive example}} \quad (13)$$

The index describing the correct extraction probability of feature extraction into the positive example is the precision rate, and its expression is:

$$\text{Precision rate}=\frac{\text{ture positive example}}{\text{ture positive example+false positive example}} \quad (14)$$

The index describing the probability that the features in the positive example are correctly extracted is recall rate, which is expressed as:

$$\text{Recall rate}=\frac{\text{ture positive example}}{\text{ture positive example+false negative example}} \quad (15)$$

The harmonic mean of precision and recall is defined as F1 value, and its expression is:

$$\text{F1}=\frac{2\times\text{ture positive example}}{2\times\text{ture positive example}-\text{false positive example+ ture negative example+false negative example}} \quad (16)$$

Among the above four indicators, there is an inverse relationship between recall rate and precision rat, that is, the higher the recall rate in the proposed method is, the lower the precision rate is. According to the prediction parameters of feature extraction, the evaluation results of multi-feature extraction of this method are obtained, as shown in Fig. 4.

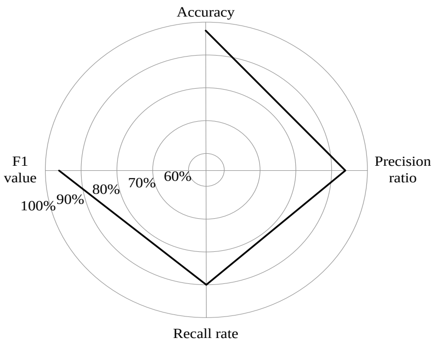

Fig. 4. Performance Analysis of Multi-feature Extraction Results.

By analyzing Fig. 4, we can see that the comprehensive performance of the four index evaluation results of the proposed method has certain advantages, indicating that the multi-feature extraction effect of this method is good.

In order to test the feature extraction efficiency of the proposed method under the condition of different sample numbers, the number of feature texts to be extracted in the experiment is set to 1000, 3000, 5000 and 10000 respectively, and the time consumed by the proposed method under different text numbers is compared. The results are shown in Table III.

TABLE III. NALYSIS OF MULTI-FEATURE EXTRACTION EFFICIENCY OF THIS METHOD

Elapsed time /s	Number of power customer information texts / piece			
	1000	3000	5000	10000
Chinese participle	6.33	18.97	54.16	97.85
Vocabulary weight determination	2.53	8.04	18.35	26.45
Affective computing	0.86	0.96	1.22	1.31

It can be seen from Table III that the time used for text emotion calculation is relatively short when the method is used to process the power customer information text. When the sample size is 10000, the running time is less than 1.31s; It takes a long time for Chinese word segmentation. When the sample size is 10000, the running time is about 97.85s, and it grows in gradient with the gradual increase of the number of power customer portrait label texts. Among them, the time consumed by this method in Chinese word segmentation and the determination of vocabulary weight fluctuates significantly, specifically in a linear rising state; The time consumed by emotional calculation fluctuates slightly. Since the determination of Chinese word segmentation and vocabulary weight is only conducted once in the initial stage, the impact of the above two steps on the subsequent feature extraction process is not significant. The experimental results show that the increase in the number of power customer portraits has no significant impact on the efficiency of multi feature extraction, indicating that the method is suitable for power grid big data environment.

D. Label Extraction Results

Different data sources such as power system registration form information, power system customer telephone system work order and power meter information collection are the data sources of label design of power customer’s portrait. These data sources generally have the characteristics of massive, multi-directional and multi-dimensional content, high dispersion, non-uniform format and so on. To extract labels from these data sources, data association analysis and preprocessing should be carried out first. For example, by analyzing the customer information, guarantee information and charge control information in the power grid customer data, we can fully understand the attribute data of relevant types of power grid customers; By analyzing the work order information of the customer's telephone system, the actual electricity charge collection information, the electricity charge collection information, etc., we can understand the customer behavior data of the power grid. Only after fully understanding the data, it is convenient to extract the characteristics. Because the relevant data of the same power customer exists in different data tables, that is, there are several data of the same customer in different data tables, in order to better describe the power customer’s portrait, it is necessary to integrate the multi-dimensional information of the power customer, and form the portrait label of the power customer based on the multi-feature extraction results of the power customer. The label structure is shown in Table IV.

TABLE IV. POWER CUSTOMER LABEL EXTRACTION RESULTS

Label name	Label details
Basic statistics of power customer’s portrait	Age
	Gender
	Industry
	Income
	Contact information
	Contact address
	Town / village
	Customer nature
Power customer power category	Important type
	High energy consumption type
	Multi power type
	Special line variant
	Electricity charge guarantee type
	Prepaid controlled
Power customer contact preferences	Store business hall
	Network business hall
	Terminal
	APP
	Telephone system
Grid sensitive type	Non sensitive
	Class A sensitivity
	Class B sensitivity
	Class C sensitivity
	Class D sensitivity

By analyzing Table IV, we can get that the labels in the relevant information of power customers can be extracted comprehensively and accurately by using the proposed method based on the multi-feature extraction results of power customers.

E. Visualization of Power Customer’s Portrait

This paper uses Tagul to realize the visualization of power customer’s portrait, imports the label of power customer’s portrait into Tagul, and sets the size of the label according to the weight of the tag. The weight of each label in each label of power customers is set to 0.2, and the product of the frequency of each label in this feature and the feature weight are taken as the weight of the label. Because there may be the same label in several features, this paper adds the weight of the same label as the total weight of the label, and makes the power customer’s portrait of "Li Boming", as shown in Fig. 5.

According to Fig. 5, it can get the portrait results of the power customer "Li Boming". According to this portrait, it can get the power customers who are more similar to the portrait of the power customer "Li Boming". On this basis, we can provide recommendations for these power customers who are more similar to "Li Boming", improve the allocation efficiency of power application resources, and bring more convenient power consumption experience and more effective power supply services to power customers.

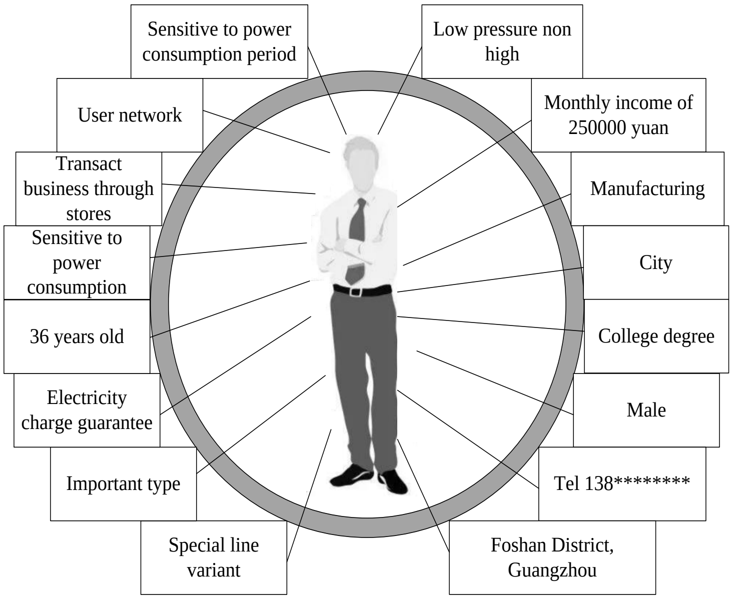
Sensitive to power consumption period
Low pressure non high
User network
Monthly income of 250000 yuan
Transact business through stores
Manufacturing
Sensitive to power consumption
City
36 years old
College degree
Electricity charge guarantee
Male
Important type
Tel 138********
Special line variant
Foshan District, Guangzhou

Fig. 5. Visualization Results of Power Portrait.

IV. Conclusion

In this paper, a multi-feature extraction method based on knowledge atlas and label extraction is proposed, and experimental research is carried out on this basis. In this paper, a multi feature extraction method based on knowledge atlas and label extraction is proposed, and experimental research is carried out on this basis. This method can realize the multi feature extraction of power users, and the calculation efficiency is high. When the sample size is less than 10000, the time for emotion calculation is less than 1.31s, the time for Chinese word segmentation is less than 97.85s, and the time for vocabulary measurement is less than 26.45s, which has certain research value. The innovation of this method lies in:

1) A construction model of power customer’s portraits based on knowledge map and label extraction is proposed, and the extracted labels are used to calculate the similarity between power customer’s portraits, which provides a research idea for personalized service recommendation and accurate power supply services to power customers through the similarity of power customer’s portraits.

2) It combs the data sources, main contents, data types and collection methods of building power customer’s portraits. By integrating multi-source data, it depicts the characteristics of power customers as people who enjoy power services in the actual power application process from a more comprehensive perspective, enriches the connotation of power customer’s portraits, presents power customer’s portraits in a visual way, and promotes the innovative development of smart power.

However, there are some limitations in the research of this paper. In the future, two main works will be carried out: expanding the sample to the data of all power customers, further enriching and constructing the group of power customer’s portraits; using PLSI, LDA, BTM and other data mining methods in topic mining to mine information in unstructured text data in various forms.

Reference

[1] H. Liang, and J. Ma, “Data-driven resource planning for virtual power plant integrating demand response customer selection and storage,” IEEE Transactions on Industrial Informatics, 2021, PP(99), pp. 1-1.

[2] F. Rahdari, N. Movahhedinia, M.R. Khayyambashi, and S. Valaee, “Qoe-aware power control and user grouping in cognitive radio ofdm-noma systems,” Computer Networks, 2021, 189(2), pp. 107906.

[3] Z. Zhao, D. Wang, H. Zhang, and H. Sang, “Joint user pairing and power allocation scheme based on transmission mode switching between noma-based maximum ratio transmission and mmse beamforming in downlink miso systems,” Mobile Information Systems, 2021(3), pp. 1-21.

[4] J. Fei, Q. Yao, M. Chen, X. Wang, and J. Fan, “The abnormal detection for network traffic of power iot based on device portrait,” Scientific Programming, 2020(9), pp. 1-9.

[5] B. Wang, E. Wang, Z. Zhu, Y. Sun, Y. Tao, and W. Wang, “An explainable sentiment prediction model based on the portraits of users sharing representative opinions in social sensors,” International Journal of Distributed Sensor Networks, 2021, 17(10), pp. 3323-3330.

[6] N. Guo, R. K. Wei, and Y. P. Shen, “Abnormal Feature Extraction Method in Large Data Environment Based on User Portrait,” Computer Simulation, 2020, 37(8), pp. 332-336.

[7] J. Chicaiza, and P. Valdiviezo-Diaz, “A comprehensive survey of knowledge graph-based recommender systems: technologies, development, and contributions,” Information (Switzerland), 2021, 12(6).

[8] J. Zhang, W. Huang, D. Ji, Y. Ren, “Globally normalized neural model for joint entity and event extraction,” Information Processing & Management, 2021, 58(5), pp. 102636.

[9] H. J. Kim, J. W. Baek, and K. Chung, “Associative knowledge graph using fuzzy clustering and min-max normalization in video contents,” IEEE Access, 2021, PP(99), pp. 1-1.

[10] J. Li, S. Liu, A. Liu, and R. Huang, “Knowledge graph construction for sofl formal specifications,” International Journal of Software Engineering and Knowledge Engineering, 2022, 32(04), pp. 605-644.

[11] J. T. Ma, J. W. Yan, t. C. Xue, and Q. Q. Ya, “Ngdcrm:a numeric graph dependency-based conflict resolution method for knowledge graph,” 2021, 27(2), pp. 10.

[12] Z. H. Han, X. S. Chen, X. M. Zeng, Y. Zhu, and M. Y. Yin, "Detecting proxy user based on communication behavior portrait," The Computer journal, 2019, 62(12), pp. 1777-1792.

[13] X. Li, K. Zhang, Q. Zhu, Y. Wang, and J. Ma, "Hybrid feature fusion learning towards chinese chemical literature word segmentation," IEEE Access, 2021, PP(99), pp. 1-1.

[14] S. Ying, W. Li, B. He, W. Wang, and Y. Wan, "Chinese segmentation of city address set based on the statistical decision tree," Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 2019, 44(2), pp. 302-309.

[15] Z. Kong, C. Yue, Y. Shi, J. Yu, and L. Xie, "Entity extraction of electrical equipment malfunction text by a hybrid natural language processing algorithm," IEEE Access, 2021, PP(99), pp. 1-1.

[16] F. Dornaika, A. Baradaaji, and Y. E. Traboulsi, "Semi-supervised classification via simultaneous label and discriminant embedding estimation," Information Sciences, 2020, 546(1).

[17] Z. Taskin, and U. Al, "Natural language processing applications in library and information science," Online Information Review, ahead-of-print 2019, (4), pp. 676-690.

[18] L. Niu, J. Cai, A. Veeraraghavan, and L. Zhang, "Zero-shot learning via category-specific visual-semantic mapping and label refinement," IEEE Transactions on Image Processing, 2019, 28(2), pp. 965-979.

[19] Z. C. Sha, Z. M. Liu, C. Ma, and J. Chen, "Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information," Applied Intelligence, 2021, 51(22).

[20] T. Sun, C. Zhang, Y. Ji, and Z. Hu, "A latent-label denoising method for relation extraction with self-directed confidence learning," Intelligent Data Analysis, 2020, 24(1), pp. 101-117.